

# Detecting and Characterizing Small Dense Bipartite-like Subgraphs by the Bipartiteness Ratio Measure

Angsheng Li\*

Chinese Academy of Sciences

Pan Peng†

Chinese Academy of Sciences

## Abstract

We motivate the problem of finding small subgraphs with small *bipartiteness (ratio)* as a variant of detecting small cyber-communities in the Web graph. The bipartiteness ratio of a subgraph  $S$ , as introduced by Trevisan [Tre09], roughly measures how close of  $S$  being a dense bipartite subgraph. We give a bicriteria approximation algorithm  $\text{SwpDB}$  such that if there exists a subset  $S$  of volume at most  $k$  and bipartiteness ratio  $\theta$ , then for any  $0 < \epsilon < 1/2$ , it finds a set  $S'$  of volume at most  $2k^{1+\epsilon}$  and bipartiteness at most  $4\sqrt{\theta/\epsilon}$ .

By combining a truncation operation, we give a local algorithm  $\text{LocDB}$ , which has asymptotically the same approximation guarantee as the algorithm  $\text{SwpDB}$  on both the volume and bipartiteness of the output set, and runs in time  $O(\epsilon^2 \theta^{-2} k^{1+\epsilon} \ln^3 k)$ , independent of the size of the graph. Our local algorithm is the first sublinear (in the size of the input graph) time algorithm with almost the same guarantee as Trevisan's spectral inequality that relates the bipartiteness of the graph to the largest eigenvalue of the (normalized) Laplacian of the graph, and runs in time slightly super linear in the size of the output set. Finally, we give a spectral characterization of the small dense bipartite-like subgraphs by using the  $k$ th largest eigenvalue of the Laplacian of the graph, which is of independent interest since most of previous spectral characterizations of combinatorial objects only use the first  $k$  smallest eigenvalues.

## 1 Introduction

Community detection and characterization has stimulated widespread interest in modern network science, which has been a very active research area due to the proliferation of very large social and technological networks over the past few years. In the literature of computer science, communities are often referred to as locally dense subgraphs in which edges are densely connected with each other while loosely connected to the outside of the subgraph. Communities convey valuable information on both the structures and dynamics of networks, and have found applications in market advertising, rumor spreading, ranking web pages and so on. For more motivations and detection methods, see recent surveys [Sch07, POM09, For10].

In this paper, we focus on the problem of searching and characterizing the *cyber-communities*, which, as argued by Kumar et al. [KRRT99], are well characterized by *dense bipartite subgraphs* due to the particular phenomenon of heavy *co-citations* among related web pages in the Web, that is, related pages are frequently referenced together. Here a dense bipartite subgraph refers to a subgraph that is sparsely connected to the outside and can be partitioned into two disjoint vertex sets  $L, R$  such that many of the possible edges between  $L$  and  $R$  are present. Since the work of Kurmar et al. [KRRT99], practitioners have proposed a large set of simple and efficient heuristic methods to extract this kind of subgraphs (eg., [KMS04, DGP09]). These heuristics are often case-by-case and experimental. On the

---

\*State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, email: angsheng@ios.ac.cn

†State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, email: pengpan@ios.ac.cn

other hand, to our knowledge, theoreticians have only studied the extreme cases of the dense bipartite subgraphs, eg., the maximum edge bicliques [Pee03], which are far from being close to the true cyber-communities detected in the Web; there is no algorithm with provable approximation or running time guarantees for a more proper measure of a set being a dense bipartite subgraph.

Let us elaborate more on how to measure a dense bipartite-like subgraph. Let  $G = (V, E)$  be an undirected graph representing a Web graph, in which an edge between two nodes indicates the existence of a hyperlink between the corresponding two web pages. (We ignore the direction of the links.) As stated above, a dense bipartite-like subgraph is a pair of disjoint vertex subsets  $L, R$  such that ‘most’ of the edges involving the vertices in  $U := L \cup R$  lie between  $L$  and  $R$ . Equivalently, we say that  $L, R$  form a dense bipartite subgraph if ‘few’ edges lie totally in  $L$  or  $R$ , or leaving  $U$  to the rest of the graph. The latter formulation turns out to be well captured by the bipartiteness ratio (shorted as bipartiteness) measure of  $L, R$ , which was introduced by Trevisan with a totally different motivation to serve as a subroutine for designing approximation algorithms for Max Cut problem [Tre09]. The bipartiteness of  $L, R$  is defined as

$$\beta(L, R) = \frac{2e(L) + 2e(R) + e(U, \bar{U})}{\text{vol}(U)},$$

where  $e(L), e(U, \bar{U})$  denote the number of edges in  $L$  and the number of edges leaving from  $U$  to the rest of the graph, respectively; and  $\text{vol}(U)$ , called the volume of  $U$ , is defined to be the sum of degree of vertices in  $U$ . Notice that the numerator involves all the edges that are *not* between  $L$  and  $R$ , and the denominator involves all the edges incident to  $L \cup R$ . It is intuitive that the smaller the bipartiteness, the more likely it behaves like a dense bipartite subgraph.

Thus, we will use the bipartiteness as a measure of a set being dense bipartite-like. We want to extract subgraphs with small bipartiteness, which corresponds to good cyber-communities. Furthermore, we are interested in finding *small* communities, which generally contains more interesting and substantial information than large communities partly due to the hierarchical organization of the community structure in networks, that is, large communities are usually consisted of small ones. Furthermore, Leskovec et al [LLDM09, LLM10] find that in many large scale networks, the sets which mostly resemble communities are of size around 100, which is rather small compared to the size of the network. There is also experimental evidence and common experience that a significant fraction of nodes in networks belong to some small communities, which is mathematically characterized as the *small community phenomenon in networks* [LP11, LP12].

In order to make our algorithm practical, we would like to design a local algorithm to extract subgraphs with small bipartiteness. A local algorithm, introduced by Spielman and Teng [ST04], is one that given as input a vertex, it only explores a small portion of the graph and finds a subgraph with good property, which has found applications in graph sparsification, solving linear equations [Spi10], and designing near-linear time algorithms [Ten10]. Local algorithms have also shown to be both effective and efficient on real network data (e.g., [LLDM09, LLB<sup>+</sup>09]).

## 1.1 Our Results

We give approximation, local algorithms and spectral characterization of the finding the small subgraphs with small bipartiteness, as we argued above, with the goal of extracting small cyber-communities. In the following, we will use the terminology of small dense bipartite-like subgraphs to indicate small subgraphs with small bipartiteness.

- We first give a bicriteria approximation algorithm for finding the small dense bipartite-like subgraph, and thus determining the *dense bipartite profile* of the graph, which is defined as

$$\beta(k) := \min_{\substack{L, R: L \cap R = \emptyset \\ \text{vol}(L \cup R) \leq k}} \beta(L, R).$$

More specifically, we give a polynomial time algorithm  $\text{SwpDB}$  such that for any  $0 < \epsilon < 1/2$ , if the graph contains a subgraph  $S$  with volume at most  $k$  and bipartiteness  $\theta$ , then it finds a subgraph  $X$  with volume at most  $2k^{1+\epsilon}$  and bipartiteness at most  $4\sqrt{\theta/\epsilon}$ .

Note that the approximation ratio does not depend on the size of the graph, since the algorithm is based on a spectral characterization of the bipartiteness of the graph given by Trevisan [Tre09] (see Lemma 1), which is analogous to the Cheeger's inequality for *conductance* (see more discussions below).

- By incorporating a truncation operation we are able to give a *local algorithm* for the dense bipartite subgraphs. We show that if the graph contains a subgraph  $S$  with volume at most  $k$  and bipartiteness at most  $\theta$ , then there exists a subgraph  $S_\theta \subseteq S$  with volume at least  $\text{vol}(S)/2$  such that if our local algorithm  $\text{LocDB}$  takes as input a vertex  $v \in S_\theta$ , then for any  $0 < \epsilon < 1/2$ , it finds a subgraph  $X$  with volume at most  $O(k^{1+\epsilon})$  and  $O(\sqrt{\theta/\epsilon})$ , with running time  $O(\epsilon^2 \theta^{-2} k^{1+\epsilon} \ln k^3)$ , independent of the size of the graph. We remark that the algorithm runs in sublinear time (in the size of the input graph, denoted as  $n$ ) when the size of the optimal set is sufficiently smaller than  $n$  and the approximation ratio of the algorithm is almost optimal in that it almost matches the guarantee of Trevisan's spectral inequality for the bipartiteness.
- Finally, as an application of the algorithm  $\text{SwpDB}$ , we give a spectral characterization of the small dense bipartite subgraph. Let  $\lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$  be the eigenvalues of the Laplacian matrix  $\mathcal{L}$  of the graph  $G$ . We showed that if  $\lambda_{n-k} \geq 2 - 2\eta$ , then there is a polynomial time algorithm such that for any  $0 < \epsilon < 1$ , it finds a subgraph with volume at most  $O(\text{vol}(G)/k^{1-\epsilon})$  and bipartiteness at most  $O(\sqrt{(\eta/\epsilon) \log_k n})$ , where  $\text{vol}(G)$  is the total degree of vertices in  $G$ . One can interpret the result as

$$\beta(\text{vol}(G)/k^{1-\epsilon}) \leq O(\sqrt{(2 - \lambda_{n-k}) \log_k n}).$$

Note that we related the  $k$ th largest eigenvalue of  $\mathcal{L}$  with some combinatorial object (in this case, the small dense bipartite subgraph), which is of independent interest as previous works mostly just use the first  $k'$  smallest eigenvalue to characterize some combinatorial objects (e.g., small set expander) in graphs (see more discusses below).

## 1.2 Our Techniques

Our approximation algorithm is based on Trevisan's spectral characterization of the bipartiteness  $\beta(G)$  of the graph, which is the minimum bipartiteness of all possible disjoint vertex subsets  $L, R$ , that is,  $\beta(G) = \beta(\text{vol}(G))$ . Recall that  $\lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$  are the eigenvalues of  $\mathcal{L}$ . Instead of working directly on  $\mathcal{L}$ , we study a closely related matrix  $M$ , which we call the *quasi-Laplacian*, that has the same spectra as  $\mathcal{L}$ . Let  $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{n-1}$  be the corresponding eigenvectors of  $M$ . Trevisan showed that if  $\lambda_{n-1} \geq 2 - 2\theta$ , then by a simple *sweeping process* over the largest eigenvector  $\mathbf{v}_{n-1}$ , we can find a pair of subsets  $X, Y$  with bipartiteness at most  $2\sqrt{\theta}$ . On the other hand, it is well known that the largest eigenvector  $\mathbf{v}_{n-1}$  can be computed fast by the power method, which starts with a "good" vector  $\mathbf{q}_0$  and iteratively multiplies it by  $M$  to obtain  $\mathbf{q}_t$ , and outputs  $\mathbf{q}_T$  by choosing proper  $T$ . Hence, the power method combined with the sweep process can find a subset with bipartiteness close to  $\beta(G)$ . However, such a method does not give a useful volume bound on the output set.

In order to find *small* dense bipartite subgraphs, we sweep each of the vector  $\mathbf{q}_t$  and characterize  $\mathbf{q}_t$  in terms of the minimum of bipartiteness of all the small sweep sets (the sets found in the sweeping process) encountered in all the  $T$  iterations. This is done by a potential function  $J(\mathbf{p}, x)$ , which has a nice convergence property that for general vector  $\mathbf{p}$  and some  $x$ ,  $J(\mathbf{p}M, x)$  can be bounded by a function of  $J(\mathbf{p}, x')$  and the bipartiteness of the some sweep set (see Lemma 2). Using this property, we show inductively that if we choose  $\mathbf{q}_0 = \chi_v$  for some vertex  $v \in V$ ,  $J(\mathbf{q}_t, x)$  can be upper bounded by a function in  $t, K$  and the minimum bipartiteness of all the sweep sets of volume at most  $K$  for all  $t \leq T$  (see Lemma 3). On the other hand, if the graph contains a small dense bipartite subgraph  $L, R$  of volume at most  $k$ , we prove that the potential function also increases quickly in terms of  $t$  and  $\beta(L, R)$  (see Lemma 4), which will lead to the conclusion that at least one of the sweep set with volume

at most  $K$  has bipartiteness “close” to  $\beta(L, R)$  by choosing proper  $K$  in terms of  $k$  and the starting vertex  $v$ .

To give local algorithms that run in time independent of the size of the graph, we need to keep the support size of the vectors  $\mathbf{q}_t$  small in each iteration. This is done by a truncation operation of a vector that only keeps the elements with large absolute vector value. Let  $\tilde{\mathbf{q}}_0 = \chi_v$  and iteratively define  $\tilde{\mathbf{q}}_t$  to be the truncation vector of  $\tilde{\mathbf{q}}_{t-1}M$ . We show that both upper bound and lower bound on  $J(\mathbf{q}_t, x)$  still approximately holds for  $J(\tilde{\mathbf{q}}_t, x)$ , and thus prove the correctness of our local algorithm which sweeps all the vectors  $\tilde{\mathbf{q}}_t$  instead of  $\mathbf{q}_t$ .

Finally, we use a simple trace lower bound to serve as the lower bound for  $J(\mathbf{q}_t, x)$  and obtain the spectral characterization of the dense bipartite profile.

### 1.3 Related Works

Our work is closely related to a line of research on the *conductance* of a set  $S$ , which is defined as

$$\phi(S) = \frac{e(S, \bar{S})}{\min\{\text{vol}(S), \text{vol}(\bar{S})\}}.$$

Kannan, Vempala and Veta [KVV04] suggest using the conductance as a measure of a set being a general community (in contrast of cyber-communities), since the smaller the conductance it, the more likely that the set is a community with dense intra-connections and sparse inter-connections. Spielman and Teng give the first local clustering algorithm to find subgraphs with small conductance by using the truncated random walk [ST04, ST08]. Anderson, Chung and Lang [ACL06], Anderson and Peres [AP09], Kwok and Lau [KL12] and Oveis Gharan and Trevisan [OT12] then give local algorithms for conductance with better approximation ratio or running time. All their local algorithms are based on the Cheeger’s inequality that relates the second smallest eigenvalue of  $\mathcal{L}$  to the conductance [AM85, Alo86, SJ89], similar to our algorithms which depend on Trevisan’s spectral inequality that relates the largest eigenvalue of  $\mathcal{L}$  to the bipartiteness.

Some works studied the small set expander graph, that is, to find small set with small conductance. This problem is of interest not only for the reason that it has applications in finding small communities, but also that it is closely related to the unique games conjecture [RS10]. Arora, Barak and Steurer [ABS10], Louis, Raghavendra, Tetali and Vempala [LRTV12], Lee, Oveis Gharan and Trevisan [LOT12], Kwok and Lau [KL12], Oveis Gharan and Trevisan [OT12] and O’Donnell and Witmer [OW12] have given spectra based approximation algorithms and characterizations of this problem. The latter three works have recently shown that for any  $0 < \epsilon < 1$ ,

$$\phi(\text{vol}(G)/k^{1-\epsilon}) \leq O(\sqrt{\lambda_k \log_k n}),$$

where  $\phi(k)$  is the *expansion profile* of  $G$  and is defined as

$$\phi(k) := \min_{S: \text{vol}(S) \leq k} \phi(S).$$

Their spectral characterization of the expansion profile as well as the Cheeger’s inequality all use the first  $k$  smallest eigenvalues of  $\mathcal{L}$ , which is comparable to our characterization of the dense bipartite profile by the  $k$ th largest eigenvalue of  $\mathcal{L}$ .

Peng [Pen12] has given a local algorithm for the dense bipartite subgraphs. His algorithm is guaranteed to output a set with volume at most  $O(k^2)$  and bipartiteness  $O(\sqrt{\theta})$ , which is worse than the approximation guarantee in our local algorithm when  $\epsilon < 1/2$  is a constant.

## 2 Preliminaries

Let  $G = (V, E)$  be an undirected weighted graph and let  $n := |V|$  and  $m := |E|$ . Let  $d(v)$  denote the weighted degree of vertex  $v$ . For any vertex subset  $S \subseteq V$ , let  $\bar{S} := V \setminus S$  denote the complementary of

$S$ . Let  $e(S)$  be the number of edges in  $S$  and define the volume of  $S$  to be the sum of degree of vertices in  $S$ , that is  $\text{vol}(S) := \sum_{v \in S} d(v)$ . Let  $\text{vol}(G) := \text{vol}(V) = 2m$ . For any two subsets  $L, R \subseteq V$ , let  $e(L, R)$  denote the number of edges between  $L$  and  $R$ . For two disjoint subsets  $L, R$ , that is,  $L \cap R = \emptyset$ , we will use  $U = (L, R)$  to denote subgraph induced on  $L$  and  $R$ , which is also called the pair subgraph. We will also use  $U$  to denote  $L \cup R$ . Given  $U = (L, R)$ , the *bipartiteness (ratio) of  $U$*  is defined as

$$\beta(L, R) := \frac{2e(L) + 2e(R) + e(U, \bar{U})}{\text{vol}(U)}.$$

The *bipartiteness of a set  $S$*  is defined to be the minimum value of  $\beta(L, R)$  over all the possible partitions  $L, R$  of  $S$ , that is,

$$\beta(S) := \min_{(L, R) \text{ partition of } S} \beta(L, R).$$

The bipartiteness of the graph  $G$  is defined as

$$\beta(G) := \beta(V) = \min_{S \subseteq V} \beta(S).$$

We are interested in finding small subgraphs with small bipartiteness. In the following, we use lower bold letters to denote vectors. Unless otherwise specified, a vector  $\mathbf{p}$  is considered to be a row vector, and  $\mathbf{p}^T$  is its transpose. For a vector  $\mathbf{p}$  on vertices, let  $\text{supp}(\mathbf{p})$  denote the support of  $\mathbf{p}$ , that is, the set of vertices on which the  $\mathbf{p}$  value is nonzero. Let  $\|\mathbf{p}\|_1$  and  $\|\mathbf{p}\|_2$  denote the  $L^1$  and  $L^2$  norm of  $\mathbf{p}$ , respectively. Let  $|\mathbf{p}|$  denotes its absolute vector, that is,  $|\mathbf{p}|(v) := |\mathbf{p}(v)|$ . For a vector  $\mathbf{p}$  and a vertex subset  $S$ , let  $\mathbf{p}(S) := \sum_{v \in S} \mathbf{p}(v)$ . For  $L, R$ , let  $\mathbf{p}(L, -R) := \sum_{v \in L} \mathbf{p}(v) - \sum_{v \in R} \mathbf{p}(v)$ . One useful observation is that for any partition  $(L, R)$  of  $S$ ,  $\mathbf{p}(L, -R) \leq |\mathbf{p}|(S)$ . Also note that there exists a partition  $(L_0, R_0)$  of  $S$  such that  $\mathbf{p}(L_0, -R_0) = |\mathbf{p}|(S)$ . Actually,  $L_0$  is the set of vertices with positive  $\mathbf{p}$  value and  $R_0$  is the set of the remaining vertices, that is,  $L_0 = \{v \in S : \mathbf{p}(v) > 0\}$  and  $R_0 = \{v \in S : \mathbf{p}(v) \leq 0\}$ .

For any vertex  $v$ , let  $\chi_v$  denote the indicator vector on  $v$ . Let  $\mathbf{1}$  denote the all 1 vector. For a set  $U = (L, R)$ , define  $\rho_U$  and  $\psi_U$  as

$$\rho_U(v) = \begin{cases} d(v)/\text{vol}(U) & \text{if } v \in L, \\ -d(v)/\text{vol}(U) & \text{if } v \in R, \\ 0 & \text{otherwise.} \end{cases} \quad \psi_U(v) = \begin{cases} \sqrt{d(v)/\text{vol}(U)} & \text{if } v \in L, \\ -\sqrt{d(v)/\text{vol}(U)} & \text{if } v \in R, \\ 0 & \text{otherwise.} \end{cases}$$

Now let  $A$  denote the adjacency matrix of the graph such that  $A_{uv}$  is the weight of edge  $u \sim v$ . Let  $D$  denote the diagonal degree matrix. Define the *random walk matrix*  $W$ , the (*normalized*) *Laplacian matrix*  $\mathcal{L}$  and the *quasi-Laplacian matrix*  $M$  of the graph  $G$  as

$$W := D^{-1}A, \mathcal{L} := I - D^{-1/2}AD^{-1/2}, M := I - D^{-1}A.$$

It is well known that these three matrices are closely related. In particular, if we will let  $\lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$  be the eigenvalues of  $\mathcal{L}$ , then  $\{1 - \lambda_i\}_{0 \leq i \leq n-1}$  and  $\{\lambda_i\}_{0 \leq i \leq n-1}$  are the eigenvalues of  $W$  and  $M$ , respectively. In this paper, we will mainly use the quasi-Laplacian  $M$  to give both algorithms and spectral characterization for the small dense bipartite subgraph problem. If we let  $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{n-1}$  be the corresponding eigenvectors of  $M$ , then we have the following spectral inequality given by Trevisan [Tre09] (see also [Pen12]).

**Lemma 1** ([Tre09]). *Let  $\beta(G)$ ,  $\lambda_{n-1}$  and  $\mathbf{v}_{n-1}$  defined as above. We have that,*

$$\beta(G) \leq \sqrt{2(2 - \lambda)}. \quad (1)$$

Furthermore, a pair subgraph  $(X, Y)$  with bipartiteness  $\sqrt{2(2 - \lambda)}$  can be found by a sweeping process over  $\mathbf{v}_{n-1}$ .

The sweeping process mentioned above is defined as follows.

**Definition 1.** (Sweep process) Given a vector  $\mathbf{p}$ , the sweep (process) over  $\mathbf{p}$  is defined by performing the following operations:

- Order the vertices so that

$$\frac{|\mathbf{p}(v_1)|}{d(v_1)} \geq \frac{|\mathbf{p}(v_2)|}{d(v_2)} \dots \geq \frac{|\mathbf{p}(v_n)|}{d(v_n)}$$

- For each  $i \leq n$ , let  $L_i(\mathbf{p}) := \{v_j : \mathbf{p}(v_j) > 0 \text{ and } j \leq i\}$ ,  $R_i(\mathbf{p}) := \{v_j : \mathbf{p}(v_j) \leq 0 \text{ and } j \leq i\}$  and  $S_i(\mathbf{p}) := (L_i(\mathbf{p}), R_i(\mathbf{p}))$ , which we call the sweep set of the first  $i$  vertices. Compute the bipartiteness of  $S_i(\mathbf{p})$ .

In Trevisan's inequality, to find the subgraph with small bipartiteness, we just need to output the sweep set with the minimum bipartiteness over all the sweep sets. Trevisan also showed the tightness (within constant factors) of inequality (1) in the sense that there exist graphs such that the two quantities in both hands of the inequality are asymptotically the same. The sweeping process as well as Trevisan's inequality are the bases of our algorithms for the small dense bipartite-like subgraphs.

We will use the following truncation operator to design local algorithms.

**Definition 2.** (Truncation operator) Given a vector  $\mathbf{p}$  and a nonnegative real number  $\xi$ , we define the  $\xi$ -truncated vector of  $\mathbf{p}$  to be:

$$[\mathbf{p}]_\xi(u) = \begin{cases} \mathbf{p}(u) & \text{if } |\mathbf{p}(u)| \geq \xi d(u) \\ 0 & \text{otherwise} \end{cases}$$

The following facts are straightforward.

**Fact 1.** For any vector  $\mathbf{p}$  and  $0 \leq \xi \leq 1$ ,

1.  $||[\mathbf{p}]_\xi| \leq |\mathbf{p}| \leq |[\mathbf{p}]_\xi| + \xi \mathbf{d}$ , where  $\mathbf{d}$  is the degree vector.
2.  $\text{vol}(\text{supp}([\mathbf{p}]_\xi)) = \sum_{v \in \text{supp}([\mathbf{p}]_\xi)} d(v) \leq \sum_{v \in \text{supp}([\mathbf{p}]_\xi)} |\mathbf{p}(v)| / \xi \leq \|\mathbf{p}\|_1 / \xi$ .

### 3 Approximation Algorithm for the Small Dense Bipartite-like Subgraphs

In this section, we first give the description of our approximation algorithm for the small dense bipartite-like subgraph, the main subroutine of which is the sweeping process over a set of vectors  $\chi_v M^t$ . We then introduce a potential function  $J(\mathbf{p}, x)$  and give both upper bound and lower bound of the potential function  $J(\chi_v M^t)$  under certain conditions, using which we are able to show the correctness of our algorithm.

#### 3.1 Description of the Algorithm and the Main Theorem

Now we describe our algorithm SwpDB (short for “sweep for dense bipartite”) for finding the small dense bipartite-like subgraphs.

SwpDB( $k, \theta, \epsilon$ )
Input: A target volume $k$ , a target bipartiteness $\theta$ , an error parameter $\epsilon < 1/2$ .
Output: A subgraph $(X, Y)$ .
<ol style="list-style-type: none"> <li>1. Let <math>T = \frac{\epsilon \ln ck}{2\theta}</math>, where <math>c</math> is some constant such that <math>c^{-\epsilon} - c^{-1} &gt; 1/2</math>. Let <math>K = 2k^{1+\epsilon}</math>.</li> <li>2. Sweep over all vectors <math>\chi_v M^t</math>, for each vertex <math>v \in V</math> and <math>t \leq T</math>, to obtain a family <math>\mathcal{F}</math> of sweep sets with volume at most <math>K</math>.</li> <li>3. Output the subgraph <math>(X, Y)</math> with the smallest bipartiteness ratio among all sets in <math>\mathcal{F}</math>.</li> </ol>

Our main theorem of this algorithm is as follows.

**Theorem 1.** *Assume that  $G$  has a set  $U = (L, R)$  such that  $\beta(L, R) \leq \theta$  and  $\text{vol}(U) \leq k$ , then for any  $0 < \epsilon < 1/2$ , the algorithm  $\text{SWPDB}(k, \theta, \epsilon)$  runs in polynomial time and finds a set  $(X, Y)$  such that  $\text{vol}(X \cup Y) \leq 2k^{1+\epsilon}$ , and  $\beta(X, Y) \leq 4\sqrt{\theta/\epsilon}$ .*

### 3.2 A Potential Function

We define a potential function  $J : [0, 2m] \rightarrow \mathbb{R}^+$ :

$$J(\mathbf{p}, x) := \max_{\substack{\mathbf{w} \in [0, 1]^n \\ \mathbf{w}(v)d(v)=x}} \sum_{v \in V} |\mathbf{p}(v)|\mathbf{w}(v).$$

Note that our potential function is similar to a potential function for bounding the convergence of  $\mathbf{p}(\frac{L+W}{2})^t$  in terms of the conductance given by Lovász and Simonovits [LS90, LS93]. Here we will use  $J(\mathbf{p}, x)$  to bound the convergence of  $\mathbf{q}M^t$  in terms of the bipartiteness of the sweep sets.

There are two useful ways to see this potential function:

- We view each edge  $u \sim v \in E$  as two directed edges  $u \rightarrow v$  and  $v \rightarrow u$ . For each directed edge  $e = u \rightarrow v$ , let  $\mathbf{p}(e) = \frac{\mathbf{p}(u)}{d(u)}$ . Order the edges so that

$$|\mathbf{p}(e_1)| \geq |\mathbf{p}(e_2)| \cdots \geq |\mathbf{p}(e_{2m})|$$

Now we can see that for an integer  $x$ ,  $J(\mathbf{p}, x) = \sum_{j=1}^x |\mathbf{p}(e_j)|$ . For other fractional  $x = \lfloor x \rfloor + r$ ,  $J(\mathbf{p}, x) = (1-r)J(\mathbf{p}, \lfloor x \rfloor) + rJ(\mathbf{p}, \lceil x \rceil)$ .

Also it is easy to see that for any directed edge set  $F$ ,  $|\mathbf{p}|(F) := \sum_{e \in F} |\mathbf{p}(e)| \leq J(\mathbf{p}, |F|)$ , since the former is a sum of  $|\mathbf{p}|$  values of one specific set of edges with  $|F|$  edges and the latter is the maximum over all such possible edge sets.

- Another way to view the potential function is to use the sweep process over  $\mathbf{p}$  as in Definition 1. By the definitions of the potential function and the sweep process, we have the following observations.
  1. For  $x = \text{vol}(S_i(\mathbf{p}))$ , then  $J(\mathbf{p}, x) = \sum_{j=1}^i |\mathbf{p}(v_j)| = |\mathbf{p}|(S_i(\mathbf{p})) = \mathbf{p}(L_i(\mathbf{p}), -R_i(\mathbf{p}))$ . And  $J(\mathbf{p}, x)$  is linear in other values of  $x$ .
  2. For any set  $S$ ,  $|\mathbf{p}|(S) \leq J(\mathbf{p}, \text{vol}(S))$ , since the former is the sum of  $|\mathbf{p}(v)|/d(v)$  values of vertices in  $S$  and the latter is the maximum sum over all sets with  $|S|$  vertices;

From both views, we can easily see that the potential function is a non-decreasing and concave function of  $x$ .

### 3.3 An Upper Bound for the Potential Function

Now we upper bound  $J(\mathbf{p}M, x)$  in terms of  $J(\mathbf{p}, x')$  and the bipartiteness of the sweep set of  $\mathbf{p}M$ .

**Lemma 2** (Convergence Lemma). *For an arbitrary vector  $\mathbf{p}$  on vertices, if  $\beta(L_i(\mathbf{p}), R_i(\mathbf{p})) \geq \Theta$ , then for  $x = \text{vol}(S_i(\mathbf{p}))$ ,*

$$J(\mathbf{p}M, x) \leq J(\mathbf{p}, x + \Theta x) + J(\mathbf{p}, x - \Theta x)$$

*Proof.* We show that for any  $U = (L, R)$ , we have that

$$\mathbf{p}M(L, -R) \leq J(\mathbf{p}, \text{vol}(U)(1 + \beta(L, R))) + J(\mathbf{p}, \text{vol}(U)(1 - \beta(L, R))) \quad (2)$$

Then the lemma follows by letting  $U = S_i(\mathbf{p}) = (L_i(\mathbf{p}), R_i(\mathbf{p}))$  and that

$$\begin{aligned} J(\mathbf{p}M, x) &= \mathbf{p}M(L_i(\mathbf{p}), -R_i(\mathbf{p})) \\ &\leq J(\mathbf{p}, x(1 + \beta(L_i(\mathbf{p}), -R_i(\mathbf{p})))) + J(\mathbf{p}, x(1 - \beta(L_i(\mathbf{p}), -R_i(\mathbf{p})))) \\ &\leq J(\mathbf{p}, x(1 + \Theta)) + J(\mathbf{p}, x(1 - \Theta)), \end{aligned}$$

where the last inequality follows from the concavity of  $J(\mathbf{p}, x)$ .

Now we show inequality (2). Let  $L_1 \rightarrow L_2$  denote the set of direct edges from  $L_1$  to  $L_2$  for two arbitrary vertex sets  $L_1$  and  $L_2$ . We have that

$$\begin{aligned}
\mathbf{p}M(L, -R) &= \mathbf{p}(I - D^{-1}A)(L, -R) \\
&= \mathbf{p}(L) - \mathbf{p}(R) - \mathbf{p}D^{-1}A(L) + \mathbf{p}D^{-1}A(R) \\
&= \sum_{v \in L} \sum_{v \rightarrow u} \frac{\mathbf{p}(v)}{d(v)} - \sum_{v \in R} \sum_{v \rightarrow u} \frac{\mathbf{p}(v)}{d(v)} - \sum_{v \in L} \sum_{u \rightarrow v} \frac{\mathbf{p}(u)}{d(u)} + \sum_{v \in R} \sum_{u \rightarrow v} \frac{\mathbf{p}(u)}{d(u)} \\
&= \sum_{e \in L \rightarrow \bar{L}} \mathbf{p}(e) - \sum_{e \in R \rightarrow \bar{R}} \mathbf{p}(e) - \sum_{e \in R \rightarrow L} \mathbf{p}(e) - \sum_{e \in \bar{U} \rightarrow L} \mathbf{p}(e) \\
&\quad + \sum_{e \in L \rightarrow R} \mathbf{p}(e) + \sum_{e \in \bar{U} \rightarrow R} \mathbf{p}(e) \\
&\leq \sum_{e \in (L \rightarrow \bar{L}) \cup (R \rightarrow \bar{R}) \cup (\bar{U} \rightarrow U)} |\mathbf{p}(e)| + \sum_{e \in (L \rightarrow R) \cup (R \rightarrow L)} |\mathbf{p}(e)| \\
&\leq J(\mathbf{p}, 2e(L, R) + 2e(U, \bar{U})) + J(\mathbf{p}, 2e(L, R)) \\
&\leq J(\mathbf{p}, \text{vol}(U) + 2e(L) + 2e(R) + e(U, \bar{U})) \\
&\quad + J(\mathbf{p}, \text{vol}(U) - 2e(L) - 2e(R) - e(U, \bar{U}))
\end{aligned}$$

where the second to last inequality follows from the fact that  $|(L \rightarrow \bar{L}) \cup (R \rightarrow \bar{R}) \cup (\bar{U} \rightarrow U)| = 2e(L, R) + 2e(U, \bar{U})$ , that  $|(L \rightarrow R) \cup (R \rightarrow L)| = 2e(L, R)$  and that  $|\mathbf{p}|(F) \leq J(\mathbf{p}, |F|)$  for an arbitrary (directed) edge set  $F$ ; and the last inequality follows from that  $J(\mathbf{p}, x)$  is non-decreasing.  $\square$

Now we can use the convergence lemma to upper bound  $J(\chi_v M^t, x)$ .

**Lemma 3.** *For any vertex  $v \in V$ , let  $\mathbf{q}_t = \chi_v M^t$ , if for all  $t \leq T$  and all sweep sets  $S_i(\mathbf{q}_t) = (L_i(\mathbf{q}_t), R_i(\mathbf{q}_t))$  of volume at most  $K$  have bipartiteness at least  $\Theta$ , that is,  $\beta(L_i(\mathbf{q}_t), R_i(\mathbf{q}_t)) \geq \Theta$ , then for any  $t \leq T$ ,*

$$J(\mathbf{q}_t, x) \leq \frac{2^t x}{K} + \sqrt{\frac{x}{d(v)}} \left(2 - \frac{\Theta^2}{4}\right)^t$$

*Proof.* The proof is by induction and is similar to the Lemma 4.2 in [OT12].

If  $t = 0$ , then the LHS is  $x/d(v)$  for  $x \leq d(v)$  and is 1 for  $x > d(v)$ , and the RHS is at least  $\sqrt{x/d(v)}$  for any  $x \in [0, 2m]$ . Thus, the lemma holds in this case.

Assume the lemma holds for  $t - 1$ . Since  $J(\mathbf{q}_t, x)$  is piecewise linear in  $x$ , and the RHS is concave, we only need to show the lemma holds for  $x = \text{vol}(S_i(\mathbf{q}_t))$  for any  $i \leq n$ .

- For  $x > K$ , the RHS is at least  $2^t$ . On the other hand, for any vector  $\mathbf{p}$ , we have

$$J(\mathbf{p}M, 2m) = \|\mathbf{p}M\|_1 = \sum_u \left| \sum_v \mathbf{p}(v) M_{vu} \right| \leq \sum_v |\mathbf{p}(v)| \sum_u |M_{vu}| \leq 2\|\mathbf{p}\|_1 = 2J(\mathbf{p}, 2m),$$

Therefore,

$$J(\mathbf{q}_t, x) \leq J(\mathbf{q}_t, 2m) \leq 2J(\mathbf{q}_{t-1}, 2m) \leq \dots \leq 2^t J(\mathbf{q}_0, 2m) = 2^t$$

So the lemma holds for  $x$  in this case.

- For  $x \leq K$ , recall that  $x = \text{vol}(S_i(\mathbf{q}_t))$ , by Lemma 2 and the induction hypothesis, we have

$$\begin{aligned}
J(\mathbf{q}_t, x) &\leq J(\mathbf{q}_{t-1}, x + x\Theta) + J(\mathbf{q}_{t-1}, x - x\Theta) \\
&\leq \frac{2 * 2^{t-1} x}{K} + \sqrt{\frac{x}{d(v)}} \left(2 - \frac{\Theta^2}{4}\right)^{t-1} (\sqrt{1 + \Theta} + \sqrt{1 - \Theta}) \\
&\leq \frac{2^t x}{K} + \sqrt{\frac{x}{d(v)}} \left(2 - \frac{\Theta^2}{4}\right)^t,
\end{aligned}$$



where the last inequality follows from that

$$\sqrt{1+\Theta} + \sqrt{1-\Theta} \leq 2 - \frac{\Theta^2}{4}.$$

This completes the proof.  $\square$

### 3.4 A Lower Bound for the Potential Function

We show that if the graph contains a pair subgraph with small bipartiteness, then we can have a good lower bound on  $J(\chi_v M^t)$  for some vertex  $v$ . The following lemma is similar to the upper bounds on the escaping probability of random walks given by Oveis Gharan and Trevisan [OT12].

**Lemma 4.** *If  $U = (L, R)$  has bipartiteness  $\beta(L, R) \leq \theta$ , then for any integer  $t > 0$ ,*

1. *there exists a vertex  $v \in U$  such that  $|\mathbf{q}_t|(U) \geq (2 - 2\theta)^t$ , where  $\mathbf{q}_t = \chi_v M^t$ ;*
2. *there exists a subset  $U^t \subseteq U$  with  $\text{vol}(U^t) \geq \text{vol}(U)/2$  satisfying that for any  $v \in U^t$ ,  $\mathbf{q}_t = \chi_v M^t$ ,*

$$J(\mathbf{q}_t, \text{vol}(U)) \geq |\mathbf{q}_t|(U) \geq \frac{1}{400}(2 - 6\theta)^t,$$

where we have assumed that  $\theta < 1/3$ .

*Proof.* 1. For the first part, we will show that

$$\rho_U M^t(L, -R) \geq (2 - 2\theta)^t. \quad (3)$$

If it holds, then by the fact that  $\rho_U M^t(L, -R) = \sum_{v \in U} \frac{d(v)}{\text{vol}(U)} \text{sgn}(v, L) \chi_v M^t(L, -R)$ , where  $\text{sgn}(v, L)$  equals 1 if  $v \in L$  and  $-1$  if  $v \in R$ , we know there exists a vertex  $v \in U$  satisfying  $\text{sgn}(v, L) \chi_v M^t(L, -R) \geq (2 - 2\theta)^t$ . Then the lemma follows from the fact that  $|\mathbf{p}|(U) \geq \max\{\mathbf{p}(L, -R), \mathbf{p}(R, -L)\}$  for any  $\mathbf{p}$ .

To show inequality (3), we note that for any  $t \geq 0$ ,

$$\rho_U M^t(L, -R) = \rho_U D^{-1/2} \mathcal{L}^t D^{1/2}(L, -R) = \psi_U \mathcal{L}^t \psi_U^T.$$

On the other hand,

$$\begin{aligned} \psi_U (2 - \mathcal{L}) \psi_U^T &= \psi_U D^{-1/2} (D + A) D^{-1/2} \psi_U^T = \sum_{u \sim v} (\psi_U(u) / \sqrt{d(u)} - \psi_U(v) / \sqrt{d(v)})^2 \\ &= \frac{4e(L) + 4e(R) + e(U, \bar{U})}{\text{vol}(U)} \leq 2\theta, \end{aligned}$$

which implies that

$$\psi_U \mathcal{L} \psi_U^T \geq 2 - 2\theta. \quad (4)$$

Now recall that  $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1} \leq 2$  are the eigenvalues of the Laplacian  $\mathcal{L}$ . Let  $\mathbf{v}'_0, \mathbf{v}'_1, \dots, \mathbf{v}'_{n-1}$  be the corresponding orthonormal eigenvectors of  $\mathcal{L}$ . If we write  $\psi_U = \sum_i \alpha_i \mathbf{v}'_i$ , then by inequality (4), we have  $\sum_i \lambda_i \alpha_i^2 \geq 2 - 2\theta$ . Therefore,

$$\psi_U \mathcal{L}^t \psi_U^T = \sum_i \lambda_i^t \alpha_i^2 \geq \left( \sum_i \lambda_i \alpha_i^2 \right)^t \geq (2 - 2\theta)^t,$$

where the second inequality follows from the fact that  $\sum_i \alpha_i^2 = \|\psi_U\|_2^2 = 1$  and the Chebyshev's sum inequality.

2. For the second part, we show that for any set  $Z = (L_Z, R_Z)$  such that  $L_Z \subseteq L$ ,  $R_Z \subseteq R$  and  $\text{vol}(Z) \geq \frac{\text{vol}(U)}{2}$ ,

$$\rho_Z M^t(L_Z, -R_Z) \geq \frac{1}{400}(2 - 6\theta)^t, \quad (5)$$

from which we know there exists at least one vertex  $v$  in  $Z$  such that

$$|\chi_v M^t|(U) \geq |\chi_v M^t|(Z) \geq \text{sgn}(v, L_Z) \chi_v M^t(L_Z, -R_Z) \geq \frac{1}{400}(2 - 6\theta)^t.$$

Then by the choice of  $Z$ , we know that the set  $U^t := \{v : |\chi_v M^t|(U) \geq \frac{1}{400}(2 - 6\theta)^t\}$  has volume at least  $\text{vol}(U)/2$  and the lemma's statement holds.

On the other hand, we have that  $\rho_Z M^t(L_Z, -R_Z) = \psi_Z M^t \psi_Z^T$  for the same reason as in the first part of the proof, so we only need to show that

$$\psi_Z M^t \psi_Z^T \geq \frac{1}{400}(2 - 6\theta)^t.$$

Let  $H = \{i : \lambda_i \geq 2 - 6\theta\}$ . For an vector  $\mathbf{p}$ , define its  $H$ -norm as  $\|\mathbf{p}\|_H := \sqrt{\sum_{i \in H} \langle \mathbf{p}, \mathbf{v}'_i \rangle^2}$ . It is straightforward to show that  $\|\cdot\|_H$  is a seminorm. Recall that  $\psi_U = \sum_i \alpha_i \mathbf{v}'_i$  and  $\sum_i \lambda_i \alpha_i^2 \geq 2 - 2\theta$ . By the definition of  $H$ -norm and that  $\|\psi_U\|_2^2 = 1$ , we have

$$\sum_i \lambda_i \alpha_i^2 \leq 2 \sum_{i \in H} \alpha_i^2 + (2 - 6\theta) \sum_{i \notin H} \alpha_i^2 = 2\|\psi_U\|_H^2 + (2 - 6\theta)(1 - \|\psi_U\|_H^2),$$

which gives that

$$\|\psi_U\|_H^2 \geq 2/3.$$

Now we write  $\psi_Z = \sum_i \beta_i \mathbf{v}'_i$ . It is easy to show that

$$\begin{aligned} \|\psi_U - \psi_Z\|_2^2 &= \sum_{v \in Z} \left( \sqrt{\frac{d(v)}{\text{vol}(Z)}} - \sqrt{\frac{d(v)}{\text{vol}(U)}} \right)^2 + \sum_{v \in U \setminus Z} \frac{d(v)}{\text{vol}(U)} \\ &= \sum_{v \in Z} d(v) \left( \frac{1}{\text{vol}(Z)} - \frac{2}{\sqrt{\text{vol}(Z)\text{vol}(U)}} + \frac{1}{\text{vol}(U)} \right) + \frac{\text{vol}(U \setminus Z)}{\text{vol}(U)} \\ &= 2 - 2\sqrt{\frac{\text{vol}(Z)}{\text{vol}(U)}} \\ &\leq 2 - \sqrt{2}, \end{aligned}$$

where the last inequality follows from our assumption that  $\text{vol}(Z) \geq \text{vol}(U)/2$ .

Hence,

$$\|\psi_U - \psi_Z\|_H \leq \|\psi_U - \psi_Z\|_2 \leq \sqrt{2 - \sqrt{2}}.$$

Then by the triangle inequality, we have

$$\|\psi_Z\|_H \geq \|\psi_U\|_H - \|\psi_U - \psi_Z\|_H \geq \sqrt{\frac{2}{3}} - \sqrt{2 - \sqrt{2}} > \frac{1}{20}.$$

Finally, we have

$$\psi_Z \mathcal{L}^t \psi_Z = \sum_i \lambda_i^t \beta_i^2 \geq (2 - 6\theta)^t \|\psi_Z\|_H^2 > \frac{1}{400}(2 - 6\theta)^2.$$

□

### 3.5 Proof of Theorem 1

Now we are ready to prove Theorem 1.

*Proof.* Clearly the algorithm `SwpDB` runs in polynomial time. Now we show the correctness of the algorithm. Let  $\Theta = 4\sqrt{\theta/\epsilon}$ . Assume on the contrary that the algorithm `SwpDB`( $k, \theta, \epsilon$ ) does not find a desired subgraph, and thus for any  $v \in V$ , and  $t \leq T = \frac{\epsilon \ln ck}{2\theta}$ , the sweep sets  $S_i(\chi_v M^t)$  of volume at most  $K = 2k^{1+\epsilon}$  have bipartiteness at least  $4\sqrt{\theta/\epsilon}$ . Then by Lemma 3, for any  $v \in V$ ,

$$\begin{aligned} J(\chi_v M^T, k) &\leq 2^T \frac{k}{k^{1+\epsilon}} + \sqrt{k} \left(2 - \frac{\Theta^2}{4}\right)^T \leq 2^T \left(\frac{1}{2k^\epsilon} + \sqrt{k} \left(1 - \frac{2\theta}{\epsilon}\right)^{\frac{\epsilon \ln ck}{2\theta}}\right) \\ &\leq 2^T \left(\frac{1}{2k^\epsilon} + \frac{c^{-1}}{k^{1/2}}\right) \\ &< 2^T (ck)^{-\epsilon}, \end{aligned}$$

where the last inequality follows from the fact that  $\epsilon < 1/2$  and that  $c^{-\epsilon} > c^{-1} + 1/2$ .

On the other hand, since  $U = (L, R)$  is subgraph such that  $\beta(L, R) \leq \theta$  and  $\text{vol}(U) \leq k$ , then by Lemma 4, we know that there exists a vertex  $u \in U$  such that,

$$J(\chi_u M^T, k) \geq (2 - 2\theta)^T \geq 2^T (1 - \theta)^{\frac{\epsilon \ln ck}{2\theta}} \geq 2^T (ck)^{-\epsilon},$$

which is a contradiction.  $\square$

## 4 A Local Algorithm for Dense Bipartite-like Subgraphs

We will use the truncated operation to give our local algorithm `LocDB` (short for “local algorithm for dense bipartite subgraph”). Note that in the algorithm we just sweep the *support* of a given vector, which is important for the computation to be local.

<code>LocDB</code> ( $v, k, \theta, \epsilon$ )
Input: A vertex $v$ , a target volume $k$ , a target bipartiteness $\theta < 1/3$ and an error parameter $\epsilon < 1/2$ .
Output: A subgraph $(X, Y)$ .
1. Let $T = \frac{\epsilon \ln c_0 k}{6\theta}$ , where $c_0$ is some constant such that $c_0^{-\epsilon} \geq 800c_0^{-1} + 1$ . Let $\xi_0 = \frac{c_0^{-\epsilon} k^{-1-\epsilon}}{800T}$ , $\xi_t = \xi_0 2^t$ . Let $\tilde{\mathbf{q}}_0 := \chi_v$ , $\mathbf{r}_0 := [\tilde{\mathbf{q}}_0]_{\xi_0}$ . Let $\mathcal{F} = \emptyset$ .
2. For each time $1 \leq t \leq T$ :
(a) Compute $\tilde{\mathbf{q}}_t := \mathbf{r}_{t-1} M$ , $\mathbf{r}_t := [\tilde{\mathbf{q}}_t]_{\xi_t}$ ;
(b) Sweep over the support of $\tilde{\mathbf{q}}_t$ and add to $\mathcal{F}$ all the sweep sets.
3. Output the subgraph $(X, Y)$ with the smallest bipartiteness ratio among all sets in $\mathcal{F}$ .

**Theorem 2.** *If there is a subset  $U = (L, R)$  of volume  $\text{vol}(U) \leq k$  and bipartiteness  $\beta(L, R) \leq \theta < 1/3$ , then there exists a subgraph  $U_\theta \subseteq U$  satisfying that  $\text{vol}(U_\theta) \geq \text{vol}(U)/2$  and that if  $v \in U_\theta$ , then for any  $0 < \epsilon < 1/2$ , the algorithm `LocDB`( $v, k, \theta, \epsilon$ ) finds a subgraph  $(X, Y)$  of volume  $O(k^{1+\epsilon})$  and bipartiteness  $O(\sqrt{\theta/\epsilon})$ . Furthermore, the running time of `LocDB` is  $O(\epsilon^2 \theta^{-2} k^{1+\epsilon} \ln^3 k)$ .*

To prove the theorem, we will use the upper bound and lower bound of the potential function  $J(\mathbf{q}_t, x)$  given in Section 3. However, to show the correctness of the local algorithm, we need to work on  $J(\tilde{\mathbf{q}}_t, x)$  instead, which can be bound by combining the following properties of the truncation operations in the algorithm.

**Proposition 1.** *For any vertex  $v$ , if  $\mathbf{q}_t = \chi_v M^t$  and  $\tilde{\mathbf{q}}_t, \mathbf{r}_t$  are as defined in the algorithm `LocDB`, then for any  $t \geq 0$ ,*

1.  $\|\tilde{\mathbf{q}}_t\|_1 \leq 2^t$ ;
2.  $|\mathbf{r}_t - \mathbf{q}_t| \leq \xi_0 t 2^t \mathbf{d}$ , where  $\mathbf{d}$  is the degree vector.

*Proof.* We prove both the inequalities by induction.

1. If  $t = 0$ , the inequality trivially holds since  $\tilde{\mathbf{q}}_0 = \chi_v$ . Now assume that the inequality holds for  $t - 1$ . Then

$$\|\tilde{\mathbf{q}}_t\|_1 = \|\mathbf{r}_{t-1}M\|_1 = \|[\tilde{\mathbf{q}}_{t-1}]_{\xi_{t-1}}M\|_1 \leq \|[\tilde{\mathbf{q}}_{t-1}]_{\xi_{t-1}}\|_1 * 2 \leq 2\|\tilde{\mathbf{q}}_{t-1}\|_1 \leq 2^t,$$

where the third inequality follows by the fact that  $\|\mathbf{p}M\|_1 \leq 2\|\mathbf{p}\|_1$  for all  $\mathbf{p}$ ; the fourth inequality follows by the definition of truncation; and the last inequality follows by the induction.

2. If  $t = 0$ , the inequality holds since  $\mathbf{q}_0 = \mathbf{r}_0 = [\mathbf{q}_0]_{\xi_0} = \chi_v$ . If  $t = 1$ , then  $\mathbf{r}_1 = [\tilde{\mathbf{q}}_1]_{\xi_1} = [\mathbf{r}_0M]_{\xi_1} = [\mathbf{q}_0M]_{\xi_1} = [\mathbf{q}_1]_{\xi_1}$ , and thus  $|\mathbf{r}_1 - \mathbf{q}_1| \leq \xi_1 \mathbf{d} = 2\xi_0 \mathbf{d}$  by the Fact 1. Now assume that the inequality holds for  $t - 1$ , that is,  $|\mathbf{r}_{t-1} - \mathbf{q}_{t-1}| \leq \xi_0(t-1)2^{t-1} \mathbf{d}$ , which is equivalent to  $|(\mathbf{r}_{t-1} - \mathbf{q}_{t-1})D^{-1}| \leq \xi_0(t-1)2^{t-1} \mathbf{1}$ , where  $\mathbf{1}$  is the all 1 vector. On the other hand,

$$\begin{aligned} |\mathbf{r}_t - \mathbf{q}_t| &= |[\mathbf{r}_{t-1}M]_{\xi_t} - \mathbf{q}_t| \leq |\mathbf{r}_{t-1}M - \mathbf{q}_t| + \xi_t \mathbf{d} = |(\mathbf{r}_{t-1} - \mathbf{q}_{t-1})D^{-1}(D - A)| + \xi_t \mathbf{d} \\ &\leq 2 * \xi_0(t-1)2^{t-1} \mathbf{d} + \xi_0 2^t \mathbf{d} \\ &= \xi_0 t 2^t \mathbf{d}, \end{aligned}$$

where the second to last inequality follows from the induction hypothesis and the fact that for any vector  $\mathbf{p}$ , if  $|\mathbf{p}| \leq c\mathbf{1}$  for some constant  $c$ , then for any vertex  $v$ ,

$$|\mathbf{p}(D - A)(v)| = \left| \sum_u \mathbf{p}(u)(D_{vu} - A_{vu}) \right| \leq \sum_u |\mathbf{p}(u)|(D_{vu} + A_{vu}) \leq 2cd(v).$$

□

Note that the second part of Proposition 1 directly implies a lower bound on  $J(\tilde{\mathbf{q}}_t, x)$ . More specifically, we have the following corollary.

**Corollary 1.** *For any set  $U$ ,  $|\tilde{\mathbf{q}}_t|(U) \geq |\mathbf{r}_t|(U) \geq |\mathbf{q}_t|(U) - \xi_0 t 2^t \text{vol}(U)$ .*

Now we give an upper bound on  $J(\tilde{\mathbf{q}}_t, x)$ .

**Lemma 5.** *For any vertex  $v$ ,  $T > 0$ ,  $\Theta < 1$ , if for any  $t \leq T$ , the sweep sets  $S_i(\tilde{\mathbf{q}}_t)$  of volume at most  $K$  have bipartiteness at least  $\Theta$ , then for any  $0 \leq t \leq T$  and  $0 \leq x \leq 2m$ ,*

$$J(\tilde{\mathbf{q}}_t, x) \leq \frac{2^t x}{K} + \sqrt{\frac{x}{d(v)}} \left(2 - \frac{\Theta^2}{4}\right)^t$$

*Proof.* We prove the lemma by combining the following observations and the proof of Lemma 3.

First we note that for any  $t \leq T$  and  $x \leq 2m$ ,  $J(\mathbf{r}_t, x) \leq J(\tilde{\mathbf{q}}_t, x)$ . This follows by the definition of the potential function. More specifically, let  $\mathbf{w} \in [0, 1]^n$  be a vector that achieves  $J(\mathbf{r}_t, x)$ , that is,  $\sum_u \mathbf{w}(u)d(u) = x$  and  $J(\mathbf{r}_t, x) = \sum_v |\mathbf{r}_t|(v)\mathbf{w}(v)$ . Then  $J(\mathbf{r}_t, x) \leq \sum_v |\tilde{\mathbf{q}}_t|(v)\mathbf{w}(v) \leq J(\tilde{\mathbf{q}}_t, x)$  since for any  $v$ ,  $|\mathbf{r}_t|(v) \leq |\tilde{\mathbf{q}}_t|(v)$ . Furthermore, by the relation between  $\tilde{\mathbf{q}}_t$  and  $\mathbf{r}_{t-1}M$ , we can always guarantee that  $S_i(\tilde{\mathbf{q}}_t) = S_i(\mathbf{r}_{t-1}M)$  for every  $i \leq n$ .

Then by the conditions given in the lemma and the convergence Lemma 2, for  $x = \text{vol}(S_i(\tilde{\mathbf{q}}_t))$ , we have

$$\begin{aligned} J(\tilde{\mathbf{q}}_t, x) = J(\mathbf{r}_{t-1}M, x) &\leq J(\mathbf{r}_{t-1}, x + \Theta x) + J(\mathbf{r}_{t-1}, x - \Theta x) \\ &\leq J(\tilde{\mathbf{q}}_{t-1}, x + \Theta x) + J(\tilde{\mathbf{q}}_{t-1}, x - \Theta x) \end{aligned} \quad (6)$$

Finally, we can use the same induction as in the proof of Lemma 3 to show that the lemma's statement holds. □

Now we are ready to prove Theorem 2.

*Proof of Theorem 2.* We first show the correctness of the local algorithm and then bound its running time.

- (Correctness.) As stated in the algorithm, we choose  $T = \frac{\epsilon \ln c_0 k}{6\theta}$ . Let  $U_\theta = U^T \subseteq U$  be the subset as described in Lemma 4, which has volume at least  $\text{vol}(U)/2$ . Now let  $v \in U_\theta$  and assume that in the algorithm  $\text{LOCALDB}(v, k, \theta, \epsilon)$ , for any  $t \leq T$ , all the sweep sets  $S_i(\tilde{\mathbf{q}}_t)$  of volume at most  $800k^{1+\epsilon}$  have bipartiteness at least  $\Theta = \sqrt{48\theta/\epsilon}$ , then by Lemma 5, we have

$$\begin{aligned} J(\tilde{\mathbf{q}}_t, \text{vol}(S)) \leq J(\tilde{\mathbf{q}}_t, k) &\leq 2^t \left( \frac{k}{800k^{1+\epsilon}} + \sqrt{k} \left( 1 - \frac{\Theta^2}{8} \right)^T \right) \\ &\leq 2^t \left( \frac{1}{800k^\epsilon} + \sqrt{k} \left( 1 - \frac{6\theta}{\epsilon} \right)^{\frac{\epsilon \ln c_0 k}{6\theta}} \right) \\ &\leq 2^T \left( \frac{1}{800k^\epsilon} + \frac{c_0}{k^{1/2}} \right) \\ &< 2^T \frac{(c_0 k)^{-\epsilon}}{800}, \end{aligned}$$

where the last inequality follows from the fact that  $\epsilon < 1/2$  and that  $c_0^{-\epsilon} \geq 800c_0^{-1} + 1$ .

On the other hand, by Lemma 4 and 1 and that  $\xi_0 T = \frac{c_0^{-\epsilon} k^{-1-\epsilon}}{800}$ , we have

$$\begin{aligned} |\tilde{\mathbf{q}}_T|(U) \geq |\chi_v M^T|U| - \xi_0 T 2^T \text{vol}(U) &\geq 2^T \left( \frac{1}{400} (1 - 3\theta)^T - \xi_0 T k \right) \\ &\geq 2^T \left( \frac{1}{400} (1 - 3\theta)^{\frac{\epsilon \ln(c_0 k)}{6\theta}} - \frac{k^{-\epsilon}}{800} \right) \\ &\geq 2^T \left( \frac{1}{400} e^{-\epsilon \ln c_0 k} - \frac{(c_0 k)^{-\epsilon}}{800} \right) \\ &= \frac{2^T (c_0 k)^{-\epsilon}}{800}, \end{aligned}$$

which is a contradiction. Therefore, there exists at least one sweep set of volume at most  $O(k^{1+\epsilon})$  and bipartiteness at most  $O(\sqrt{\theta/\epsilon})$ .

- (Running time.) We first bound the time required in each iteration. For any  $t \leq T$ , instead of perform the dense vector multiplication to compute  $\tilde{\mathbf{q}}_t$ , we keep record of the support of  $\mathbf{r}_t$ , which has volume at most  $\|\tilde{\mathbf{q}}_t\|_1/\xi_t \leq 2^t/(\xi_0 2^t) = \xi_0^{-1}$ . By definition, both the volume and the computational time of  $\tilde{\mathbf{q}}_{t+1}$  are proportional to  $\text{vol}(\text{supp}(\mathbf{r}_t))$ , which is at most  $\xi_0^{-1}$  by the property of truncation operation.

During the sweep process, we only need to sweep the vertices in  $\text{supp}(\mathbf{r}_t)$ . Sorting these vertices requires time  $O(|\text{supp}(\mathbf{r}_t)| \ln |\text{supp}(\mathbf{r}_t)|) \leq O(\text{vol}(\text{supp}(\mathbf{r}_t)) \ln \text{vol}(\text{supp}(\mathbf{r}_t)))$ . Computing the bipartiteness of the sweep sets requires time  $O(\text{vol}(\text{supp}(\mathbf{r}_t)))$ . Therefore, in a single iteration, the computation takes time  $O(\text{vol}(\text{supp}(\mathbf{r}_t)) + \text{vol}(\text{supp}(\mathbf{r}_t)) \ln \text{vol}(\text{supp}(\mathbf{r}_t))) = O(\xi_0^{-1} + \xi_0^{-1} \ln \xi_0^{-1}) = O(\xi_0^{-1} \ln \xi_0^{-1})$ .

Since the algorithm takes  $T$  iterations, the total running time is thus bounded by  $O(T \xi_0^{-1} \ln \xi_0^{-1}) = O(\epsilon^2 k^{1+\epsilon} \ln^3 k / \theta^2)$ .

□

## 5 Spectral Characterization of the Small Dense Bipartite-like Subgraphs

Recall that  $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$  are the eigenvalues of its (normalized) Laplacian (and also  $M$ ).

**Theorem 3.** If  $\lambda_{n-k} \geq 2 - 2\eta$ , then there is a polynomial time algorithm such that for any  $0 < \epsilon < 1$ , it finds a subset  $(X, Y)$  of volume at most  $O(\text{vol}(G)/k^{1-\epsilon})$  and bipartiteness  $O(\sqrt{16(\eta/\epsilon) \log_k n})$ .

*Proof.* Given  $k, \eta, \epsilon$ , we set  $T = \frac{\epsilon \ln k}{2\eta}$ ,  $K = \frac{\text{vol}(G)}{0.5k^{1-\epsilon}}$ , and run the step 2 and 3 of the algorithm SWPDB to find a subgraph, which clearly runs in polynomial time. Assume that during this process, all the sweep sets  $S_i(\chi_v M^t)$  of volume at most  $K$  have bipartiteness  $\Theta = \sqrt{16(\eta/\epsilon) \log_k n}$ , for any  $v \in V$  and  $t \leq T$ . Then, by Lemma 3, we have that for any  $v \in V$ ,

$$\chi_v M^T \chi_v^T \leq J(\chi_v M_T, d(v)) \leq 2^T \frac{d(v)}{K} + \left(2 - \frac{\Theta^2}{4}\right)^T.$$

Therefore,

$$\begin{aligned} \sum_{v \in V} \chi_v M^T \chi_v^T &\leq 2^T \left( \frac{\text{vol}(G)}{K} + n \left(1 - \frac{\Theta^2}{8}\right)^T \right) \\ &= 2^T \left( 0.5k^{1-\epsilon} + n \left(1 - \frac{2\eta \log_k n}{\epsilon}\right)^{\frac{\epsilon \ln k}{2\eta}} \right) \\ &\leq 2^T (0.5k^{1-\epsilon} + 1) \\ &< 2^T k^{1-\epsilon} \end{aligned}$$

On the other hand, by the trace formula,

$$\sum_{v \in V} \chi_v M^T \chi_v^T = \text{Tr}(M^T) = \sum_{i=1}^n \lambda_i^t \geq k(2 - 2\eta)^T = 2^T k(1 - \eta)^{\frac{\epsilon \ln k}{2\eta}} \geq 2^T k^{1-\epsilon},$$

which is a contradiction.  $\square$

### Acknowledgements.

The research is partially supported by NSFC distinguished young investigator award number 60325206, and its matching fund from the Hundred-Talent Program of the Chinese Academy of Sciences. Both authors are partially supported by the Grand Project “Network Algorithms and Digital Information” of the Institute of software, Chinese Academy of Sciences.

### References

- [Alo86] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.
- [AM85] N. Alon and V.D. Milman.  $\lambda_1 \text{ sub}_i 1/\text{sub}_i$ , isoperimetric inequalities for graphs, and super-concentrators. *Journal of Combinatorial Theory, Series B*, 38(1):73–88, 1985.
- [ACL06] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.
- [AP09] Reid Andersen and Yuval Peres. Finding sparse cuts locally using evolving sets. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC ’09, pages 235–244, New York, NY, USA, 2009. ACM.
- [ABS10] S. Arora, B. Barak, and D. Steurer. Subexponential algorithms for unique games and related problems. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 563–572. IEEE, 2010.
- [DGP09] Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense implicit communities in the web graph. *ACM Transactions on the Web (TWEB)*, 3(2):7, 2009.

- [For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486, 2010.
- [KVV04] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.
- [KMS04] R. Kumar, U. Mahadevan, and D. Sivakumar. A graph-theoretic approach to extract story-lines from search results. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 216–225. ACM, 2004.
- [KRRT99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the eighth international conference on World Wide Web, WWW '99*, pages 1481–1493, New York, NY, USA, 1999. Elsevier North-Holland, Inc.
- [KL12] T.C. Kwok and L.C. Lau. Finding small sparse cuts locally by random walk. *Arxiv preprint arXiv:1204.4666*, 2012.
- [LOT12] J.R. Lee, S. Oveis Gharan, and L. Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities. *Proceedings of the 44th annual ACM symposium on Theory of computing*, 2012.
- [LLDM09] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.
- [LLM10] Jure Leskovec, Kevin J. Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 631–640, New York, NY, USA, 2010. ACM.
- [LP11] Angsheng Li and Pan Peng. Communities structures in classical network models. *Internet Mathematics*, 7(2):81–106, 2011.
- [LP12] Angsheng Li and Pan Peng. The small-community phenomenon in networks. *Mathematical Structures in Computer Science*, 22:373–407, 2012.
- [LLB<sup>+</sup>09] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. Iso-RankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, June 2009.
- [LRTV12] A. Louis, P. Raghavendra, P. Tetali, and S. Vempala. Many sparse cuts via higher eigenvalues. *Proceedings of the 44th annual ACM symposium on Theory of computing*, 2012.
- [LS90] L. Lovász and M. Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Foundations of Computer Science, 1990. Proceedings., 31st Annual Symposium on*, pages 346–354. IEEE, 1990.
- [LS93] L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.
- [OW12] R. O'Donnell and D. Witmer. Improved small-set expansion from higher eigenvalues. *Arxiv preprint arXiv:1204.4688*, 2012.
- [OT12] Shayan Oveis Gharan and Luca Trevisan. Approximating the expansion profile and almost optimal local graph clustering. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS '12*, 2012.
- [Pee03] R. Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 131(3):651–654, 2003.
- [Pen12] Pan Peng. A local algorithm for finding dense bipartite-like subgraphs. In *18th International Computing and Combinatorics Conference, COCOON '12*, 2012.
- [POM09] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56:1082–1097, 2009.

- [RS10] P. Raghavendra and D. Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 755–764. ACM, 2010.
- [Sch07] S.E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.
- [SJ89] A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing markov chains. *Information and Computation*, 82(1):93–133, 1989.
- [ST04] D.A. Spielman and S.H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2004.
- [ST08] D.A. Spielman and S.H. Teng. A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning. *Arxiv preprint arXiv:0809.3232*, 2008.
- [Spi10] Daniel A. Spielman. Algorithms, graph theory, and linear equations. IV:2698–2722, 2010.
- [Ten10] Shang-Hua Teng. The laplacian paradigm: Emerging algorithms for massive graphs. In *Theory and Applications of Models of Computation*, volume 6108 of *Lecture Notes in Computer Science*, pages 2–14. Springer Berlin / Heidelberg, 2010.
- [Tre09] Luca Trevisan. Max cut and the smallest eigenvalue. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC ’09, pages 263–272, New York, NY, USA, 2009. ACM.